Sampling Variability of Performance Assessments

Richard J. Shavelson; Gail P. Baxter; Xiaohong Gao


Stable URL:
http://links.jstor.org/sici?sici=0022-0655%28199323%2930%3A3%3C215%3ASVOPA%3E2.0.CO%3B2-M

Journal of Educational Measurement is currently published by National Council on Measurement in Education.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ncme.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.
Sampling Variability of Performance Assessments

Richard J. Shavelson, Gail P. Baxter, and Xiaohong Gao
University of California, Santa Barbara

In this article, performance assessments are cast within a sampling framework. More specifically, a performance assessment is viewed as a sample of student performance drawn from a complex universe defined by a combination of all possible tasks, occasions, raters, and measurement methods. Using generalizability theory, we present evidence bearing on the generalizability and convergent validity of performance assessments sampled from a range of measurement facets and measurement methods. Results at both the individual and school level indicate that task-sampling variability is the major source of measurement error. Large numbers of tasks are needed to get a reliable measure of mathematics and science achievement at the elementary level. With respect to convergent validity, results suggest that methods do not converge. Students' performance scores, then, are dependent on both the task and method sampled.

Performance assessments have become political instruments for educational reform in America (Bush, 1991; see also Jaeger, 1992; Shavelson, Baxter, & Pine, 1992; Shavelson, Carey, & Webb, 1990). At state and national levels (e.g., California State Department of Education, 1989; National Assessment of Educational Progress, 1987), a wide range of assessments including open-ended mathematics questions (Pandy, 1991), language arts writing samples and portfolios (Candell & Ercikan, 1992; Vermont Department of Education, 1991), and hands-on science investigations (Baron, 1990; Camplin, 1989) is being investigated. The intent is to develop measures of student achievement that focus on student ability to apply their conceptual understanding and problem-solving skills in novel situations. If assessment systems focus on "higher-order thinking," the reasoning goes, curriculum and teaching can be changed (e.g., Resnick & Resnick, 1992), and ultimately the bottom line—achievement—will be improved.

Heretofore several factors have mediated against the use of performance-based assessments in large-scale testing: cost, time, reliance on complex human judgment of questionable reliability, and lack of convergent or discriminant validity. Judging from current reform, the first two (cost and time) are no

This research was funded by grants (SPA-8751511 and TPE-9055443) from the National Science Foundation, the University of California President's Office (President's Grant for School Improvement), and the U.S. Department of Education through the National Center for Research on Evaluation, Standards, and Student Testing (No. R117G10027). We wish to thank Dr. Dale Carlson of the California Assessment Program for providing the 1990 Science Performance Assessment data. Opinions expressed reflect those of the authors and do not necessarily reflect those of the granting agencies.
longer viewed as barriers, at least for the time being (but see Shavelson, Baxter, & Pine, 1992). States such as California, Connecticut, Maryland, and Vermont are experimenting with performance assessments that are clearly more costly than multiple-choice achievement tests. The second two (reliability and validity) have, to date, received little attention. The purpose of this article is to present empirical evidence on some aspects of the technical qualities of performance assessments in elementary mathematics and science.

We view a performance assessment as a concrete, goal-oriented task (e.g., discover the contents of a “mystery box” by connecting an electric circuit to it) performed by a student on a particular occasion (e.g., sometime in the spring) and evaluated by an expert rater who takes into account the process of accomplishing the task as well as the final product. The method of presenting the task might be pencil and paper, such as an open-ended mathematics problem (e.g., Pandy, 1991); or computer, such as a simulation of a science investigation (e.g., Pine, Baxter, & Shavelson, 1991); or laboratory equipment, with experts rating performance either in real time observation or from students’ lab notebooks (e.g., Baxter, Shavelson, Goldman, & Pine, 1992).

More specifically, we view a performance assessment as a sample of student performance drawn from a complex universe defined by a combination of all possible tasks, occasions, raters, and measurement methods. We view the task facet to be representative of the content in a subject-matter domain. The occasion facet includes all possible occasions on which a decision maker would be equally willing to accept a score on the performance assessment. We view the rater facet as including all possible individuals who could be trained to score performance reliably. These three facets are, traditionally, thought of as sources of unreliability in a measurement.

In addition, we incorporate a method facet into our definition of the universe of generalization. This formulation moves us beyond reliability into a sampling theory of validity (cf. Kane, 1982). Specifically, we view the method facet to be all possible methods (e.g., short answer, computer simulation) that a decision maker would be equally willing to interpret as bearing on student achievement.

Specification of the task domain is especially critical in measuring achievement in a subject matter: Based on performance on a sample of tasks, to what domain does the decision maker generalize? One possible way to link tasks to the broader domain is suggested by Baxter, Shavelson, Herman, Brown, and Valadez (1993; see also Shavelson, Gao, & Baxter, 1992). They linked curricular goals as expressed in the California State Mathematics Framework (California State Department of Education, 1985, 1987) with teaching activities commonly used by teachers in the California Mathematics Project and translated a sample of these activities into assessments. To be used as an assessment, a goal was set for each activity. For example, ask the student to: (a) Find a problem to be solved with the activity, (b) establish criteria by which he or she would know when the problem was successfully solved, or (c) translate among alternative symbolic representations, recognizing their equivalence. This sample of activities was then translated into assessments through an iterative process of development, tryout, modification, and tryout.
Student performance estimates may vary across a sample of assessment tasks, raters, occasions, or methods. When performance varies substantially from one task sample to another, or from one occasion sample to another, or from one rater sample to another, we speak of measurement error due to sampling variability. When performance varies from one measurement method (e.g., observed performance, computer simulation, short-answer question) to another, we speak of lack of convergent validity due to method-sampling variability.

Once conceived as a sample of performance from a complex universe, the statistical framework of generalizability (G) theory can be brought to bear on the technical quality of performance-assessment scores (cf. Cronbach, Gleser, Nanda, & Rajaratnam, 1972; see also Brennan, 1991; Kane, 1982; Shavelson, Webb, & Rowley, 1989). From the G theory perspective, an assessment score or profile is but one of many possible samples from a large domain of assessments defined by the particular task, occasion, rater, measurement method (etc.). The theory focuses on the magnitude of sampling variability due to tasks, raters, and so forth, and their combinations, providing estimates of the magnitude of measurement error in the form of variance components. In addition, it provides a summary coefficient reflecting the “reliability” of generalizing from a sample score or profile to the much larger domain of interest. This coefficient is called a generalizability coefficient in G theory, recognizing that generalization may be across different facets, depending on how a performance assessment is used. The theory also can be used to estimate the magnitude of variability among scores due to method sampling, thereby providing an index of the degree to which alternative measurement methods converge (cf. Kane, 1982).

From a generalizability perspective, sampling variability due to raters, for example, speaks to a traditional concern about the viability of performance assessments—namely, interrater reliability (cf. Fitzpatrick & Morrison, 1971). Sampling variability due to tasks speaks to the complexity of the subject-matter domain for students. Traditionally, task sampling has been thought of as related to internal consistency reliability. One goal of test developers has been to make “items” homogeneous to increase reliability. Within the sampling framework, task-sampling variability is dealt with not by homogenizing the tasks but by increasing sample size from the subject-matter domain of interest (cf. Shavelson, Gao, & Baxter, 1992). Sampling variability due to occasions corresponds to the classical notion of retest reliability. From a sampling perspective, it reminds us that decision makers are willing to generalize a student’s performance on one particular occasion to many possible occasions. Finally, sampling variability due to measurement method bears on convergent validity (cf. Kane, 1982). Large method sampling variability indicates that measurement methods do not converge, as has commonly been assumed in arguing for the cost efficiency of multiple-choice testing.

Initially, technical evaluation of performance assessments focused primarily on the impact of rater sampling. With the complexity of performance assessments, the concern was that raters would be inconsistent in their evaluations.
More recently, task-sampling variability—inconsistencies in performance across tasks—has been of concern (Dunbar, Koretz, & Hoover, 1991). The findings are remarkably consistent across very diverse studies such as writing, mathematics, and science achievement of elementary students (Baxter et al., 1993; Dunbar et al., 1991; Shavelson, Baxter, & Pine, 1991) and job performance of military personnel (Shavelson, Mayberry, Li, & Webb, 1990; Wigdor & Green, 1991). Interrater reliability is not a problem, but task-sampling variability is. Large numbers of tasks are needed to get a generalizable measure.

As our sampling framework suggests, defining the universe of generalization solely in terms of tasks and/or raters is limited. With complex performance measures, a student’s achievement score may be impacted by several sources of sampling variability. Some are associated with generalizability (task, rater, and occasion sampling), and others are associated with convergent validity (method sampling). It therefore becomes important to estimate, simultaneously, as many potential sources of error—task, rater, occasion, and their interactions—and as many potential sources of method variation—methods and their interactions—as are feasible.

This article, then, presents evidence on the generalizability and convergent validity of performance assessments using data from G studies that sampled a wide range of measurement facets and measurement methods. The data were taken from studies in elementary science and mathematics. Specifically, we draw on three studies: a science assessment study funded by the National Science Foundation (Science), a mathematics assessment study funded by the Office of the President of the University of California (Math), and a science assessment conducted by the California Assessment Program (CAP-Science). These studies were chosen because they provide concrete examples of the impact of various combinations of facets and/or measurement methods on the generalizability of students’ performance scores (Brennan, 1991; Kane, 1982). Further, they illustrate the remarkable consistency of variance component estimates at the individual and school level.

**Data Sets and Analyses**

Table 1 provides an overview of the data sets used, the questions asked of the data, and the G-study designs. Table 2 presents the formulas for determining generalizability and validity coefficients.

**Science**

A team of researchers and scientists from the University of California, Santa Barbara, and the California Institute of Technology collaborated in developing and evaluating four alternative measurement methods: (a) expert observations of student performance on the hands-on tasks; (b) notebooks, in lieu of observers, in which students record the procedures they used in conducting the hands-on investigation; (c) computer simulations of tasks in which students manipulate icons on a Macintosh; and (d) short-answer problems where students answer questions about planning, analyzing, or interpreting the tasks (Shavelson et al., 1991).
### TABLE 1
Overview of Data Sets, Research Questions, and Designs

<table>
<thead>
<tr>
<th>Data Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
</tr>
<tr>
<td>Math</td>
</tr>
<tr>
<td>CAP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>( n_P )</th>
<th>Research Questions</th>
<th>Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>26</td>
<td>What is the relative impact of sampling due to raters ( r ), tasks ( t ), and occasions ( o ) ?</td>
<td>( pxrxtx0 )</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>What is the relative impact of sampling due to raters and tasks, and how does this compare to findings in math?</td>
<td>( pxrxt )</td>
</tr>
<tr>
<td></td>
<td>186</td>
<td>What is the relative impact of sampling due to tasks and methods ( m )?</td>
<td>( ptxxm )</td>
</tr>
</tbody>
</table>

What is the convergent validity of multiple measurement methods?

| Math      | 105      | What is the relative impact of sampling due to raters and tasks, and how does this compare to findings in science? | \( pxrxt \) |

| CAP       | 120      | What is the relative impact of sampling due to raters and tasks, and how does this compare to findings in math? | \( pxrxt \) |

| (\( n_{ps} = 8 \); \( n_s = 15 \)) | 120 | What is the relative impact of sampling due to persons \( p \), raters \( r \), and tasks \( t \) in measuring school-level achievement? | \( p:rxrxt \) |

One hundred and eighty-six fifth- and sixth-grade students completed each of three science tasks: Paper Towels—conduct an investigation with laboratory equipment to determine which of three paper towels holds, soaks up, or absorbs the most/least water; Electric Mysteries—use batteries, bulbs, and wires to determine the contents of six black boxes from a list of five possible alternatives (bulb, battery and bulb, wire, two batteries, or nothing); Bugs—conduct a series of experiments to determine sow bugs’ preferences for various environments (damp vs. dry, light vs. dark).

Scoring focused on both students’ procedures and conclusions, with one exception: Short-answer questions were scored right or wrong. For the Electric Mysteries task, both the circuit used to reach a conclusion about the contents of a box and the accuracy of the conclusion were taken into account when scoring. For the Paper Towels and Bugs tasks, both the procedures students used to carry out the tasks as well as their conclusions were taken into account (Baxter et al., 1992). The maximum score for each task was six points.
TABLE 2
(a) Equations for Relative Generalizability Coefficients

<table>
<thead>
<tr>
<th>Design</th>
<th>Relative G Coefficient ($\rho^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>px r x t x o*</td>
<td>$\frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2}$</td>
</tr>
<tr>
<td>px r x t</td>
<td>$\frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{pt}^2}{n_t} + \frac{\sigma_{pro}^2}{n_r n_t}}$</td>
</tr>
<tr>
<td>px t x m</td>
<td>$\frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pt}^2}{n_t} + \frac{\sigma_{pm}^2}{n_t} + \frac{\sigma_{pmt,e}^2}{n_t}}$</td>
</tr>
<tr>
<td>p: x r x t **</td>
<td>$\frac{\sigma_s^2}{\sigma_s^2 + \sigma_\delta^2}$</td>
</tr>
</tbody>
</table>

* $\sigma_\delta^2 = \frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{pt}^2}{n_t} + \frac{\sigma_{pro}^2}{n_r n_t} + \frac{\sigma_{pro}^2}{n_i} + \frac{\sigma_{pro}^2}{n_i^n_o} + \frac{\sigma_{pro}^2}{n_i^n_o} + \frac{\sigma_{pro}^2}{n_i^n_o} + \frac{\sigma_{pro}^2}{n_i^n_o}$

** $\sigma_\delta^2 = \frac{\sigma_{pr}^2}{n_p} + \frac{\sigma_{pt}^2}{n_t} + \frac{\sigma_{p,t}^2}{n_p n_t} + \frac{\sigma_{p,s}^2}{n_p n_t} + \frac{\sigma_{p,s}^2}{n_p n_t} + \frac{\sigma_{p,s}^2}{n_p n_t} + \frac{\sigma_{p,s}^2}{n_p n_t} + \frac{\sigma_{p,s}^2}{n_p n_t}$

All tasks were represented by all methods with the exception of the Paper Towels task, which could not be adequately simulated. All students ($n = 186$) were tested on all tasks and all methods. Among those 186 students, a sample of 26 students was administered each observed task and corresponding notebook on two occasions (Ruiz-Primo, Baxter, & Shavelson, 1993). In addition, two raters scored a sample of 48 students’ Paper Towels and Bugs notebooks.

For the purposes of this article, we draw three examples from this study (see Table 1). The first example, a person x rater x task x occasion G study, examined the relative contribution of raters, tasks, occasions, and their interactions to the generalizability of students’ performance scores in elementary science. The second G study examined sampling variability in a person x rater x task design with the purpose of comparing results across the Science, Math, and CAP-Science studies (Table 1). Finally, a person x task x method G study examined the relative contributions made by tasks and methods to sampling variability (cf. Kane, 1982). Table 2 presents the variance components that enter into the generalizability coefficient for each of these G studies.

Math

Teachers and researchers at the University of California, Santa Barbara, developed and evaluated mathematics performance assessments that were
Sampling Variability

<table>
<thead>
<tr>
<th>Design</th>
<th>Absolute G Coefficient (Φ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(b) Equations for Absolute Generalizability Coefficients</td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\text{p x r x t x o}^* & \quad \frac{\sigma_p^2}{\sigma_p^2 + \sigma_h^2} \\
\text{p x r x t} & \quad \frac{\sigma_p^2}{\sigma_p^2 + \sigma_r^2 + \sigma_t^2 + \sigma_{pr}^2 + \frac{\sigma_{rt}^2}{n_t} + \frac{\sigma_{rt, e}^2}{n_r n_t}} \\
\text{p: s x r x t}^{**} & \quad \frac{\sigma_s^2}{\sigma_s^2 + \sigma_h^2} \\
\end{align*}
\]

\[
\begin{align*}
\sigma_h^2 &= \frac{\sigma_r^2}{n_r} + \frac{\sigma_t^2}{n_t} + \frac{\sigma_p^2}{n_o} + \frac{\sigma_{pr}^2}{n_r n_t} + \frac{\sigma_{rt}^2}{n_r n_t} + \frac{\sigma_{rt, e}^2}{n_r n_t} + \frac{\sigma_{rt, o}^2}{n_r n_t} + \frac{\sigma_{rt, o, e}^2}{n_r n_t n_o} \\
&+ \frac{\sigma_{pr}^2}{n_r n_t} + \frac{\sigma_{pr, o}^2}{n_r n_t n_o} + \frac{\sigma_{pr, o, e}^2}{n_r n_t n_o} + \frac{\sigma_{pr, o, l}^2}{n_r n_t n_o} \\
\sigma_h^2 &= \frac{\sigma_r^2}{n_r} + \frac{\sigma_t^2}{n_t} + \frac{\sigma_p^2}{n_p} + \frac{\sigma_{pr}^2}{n_r n_t} + \frac{\sigma_{rt}^2}{n_r n_t} + \frac{\sigma_{rt, e}^2}{n_r n_t} + \frac{\sigma_{rt, o}^2}{n_r n_t} + \frac{\sigma_{rt, o, e}^2}{n_r n_t n_o} \\
&+ \frac{\sigma_{pr}^2}{n_r n_t} + \frac{\sigma_{pr, o}^2}{n_r n_t n_o} + \frac{\sigma_{pr, o, e}^2}{n_r n_t n_o} + \frac{\sigma_{pr, o, l}^2}{n_r n_t n_o} \\
\end{align*}
\]

closely aligned with hands-on instructional activities. Quality hands-on mathematics instructional activities were translated into performance assessments in the domains of measurement and place value.

One hundred and five sixth-grade students were administered the performance assessments on an individual basis. In general, the tasks confronting students were holistic in nature, involved problem solving in concrete situations with the use of manipulatives, and asked students to represent their solutions in various symbolic forms (e.g., written, graphic). Of particular interest here are the three tasks where students were asked to respond in writing. One task asked students to imagine they were talking on the telephone to a friend and to describe a green, 1 in. x 5 in. rectangular object such that the friend could draw a picture of it. A second task asked students to justify their choice of a fence perimeter for a dog run to enclose 24 sq yds (i.e., 1 x 24, 2 x 12, 3 x 8, 4 x 6). A third task asked students to compare two different representations of the sum of five numbers and explain why they were the same or different (Baxter et al., 1993).

All three tasks were scored by two raters using a 6-point holistic scoring rubric (e.g., 1 = off track, 4 = acceptable, 6 = outstanding) developed as part
of this study (cf. California State Department of Education, 1989). Here we present data from a person × rater × task G study and compare the findings with assessments in science (see Tables 1 and 2).

**CAP**

The California Assessment Program (CAP) conducted a voluntary statewide science assessment in 1989–1990 with approximately 600 schools. Students were posed five independent tasks. More specifically, students rotated through a series of five self-contained stations at timed intervals (about 15 mins.). At one station, students were asked to complete a problem solving task (determine which of these materials may serve as a conductor). At the next station, students were asked to develop a classification system for leaves and then to explain any adjustments necessary to include a new mystery leaf in the system. At yet another, students were asked to conduct tests with rocks and then use the results to determine the identity of an unknown rock. At the fourth station, students were asked to estimate and measure various characteristics of water (e.g., temperature, volume). And at the fifth station, students were asked to conduct a series of tests on samples of lake water to discover why fish are dying (e.g., is the water too acidic?). At each station, students were provided with the necessary materials and asked to respond to a series of questions in a specified format (e.g., fill in a table).

A predetermined scoring rubric developed by teams of teachers in California was used to evaluate the quality of students’ written responses (California State Department of Education, 1990) to each of the tasks. Each rubric was used to score performance on a scale from 0 to 4 (0 = no attempt, 1 = serious flaws, 2 = satisfactory, 3 = competent, 4 = outstanding). All tasks were scored by three raters. For our purposes, we report results of two G studies. The first is a person × rater × task G study carried out for comparison with the Science and Math studies (Table 1 and Table 2). The second is a person:school × rater × task design based on data from a random sample of eight students within each of a random sample of 15 schools scored by three raters using the CAP designed scoring rubric (see Tables 1 and 2).

**Results and Discussion**

We examined the sampling variability and generalizability of performance assessments at both the individual and school level. Then we examined method-sampling variability (convergent validity) across several methods of task presentation. Each of the studies presented approximates our conception of a sampling framework that includes raters, tasks, occasions, and measurement methods in its definition of performance assessments.

For each study, variance component estimates and generalizability coefficients are presented. Generalizability coefficients are calculated for both relative decisions—rank ordering students (or schools)—and absolute decisions—describing the level of performance. We speak of the former as the *relative G* coefficient and the latter as the *absolute G* coefficient. To aid in comparing the magnitude of the variance components, we present the percent
Sampling Variability of total variability accounted for by each variance component (Shavelson & Webb, 1991).

Generalizability

Sampling variability of performance assessments was examined in a series of G studies. Using the Science data, we carried out a person × rater × task × occasion G study to examine the sources of measurement error from a design that comes closest to our notion of a full model \((p \times r \times t \times o \times m)\). To examine the consistency of variance component estimates across subject matter domains, a series of person × rater × task G studies using the Science, Math, and CAP-Science data were carried out. Finally, we examined the CAP data in a person:school × rater × task design asking if variance component estimates and generalizability coefficients are similar at the school and individual level.

**Individual level.** A person × rater × task × occasion G study was carried out using the Science data (Table 3). Two raters scored the notebook performance of 26 students who completed two tasks (Bugs and Paper Towels) on two occasions (May and October).

The major source of measurement error was due to the person × task × occasion interaction (59% of the total variability). Some students performed the Paper Towels task better than the Bugs task on one occasion but performed the Bugs task more successfully than the Paper Towels task on a different occasion; vice versa for other students. The second largest source of error

### TABLE 3

Variance Component Estimates for the Person x Rater x Task x Occasion G Study Using the Science Data

<table>
<thead>
<tr>
<th>Source of Variability</th>
<th>Estimated Variance Component</th>
<th>Percent Total Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person (p)</td>
<td>0.07</td>
<td>4</td>
</tr>
<tr>
<td>Rater (r)</td>
<td>0.00^a</td>
<td>0</td>
</tr>
<tr>
<td>Task (t)</td>
<td>0.00^a</td>
<td>0</td>
</tr>
<tr>
<td>Occasion (o)</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>pr</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>pt</td>
<td>0.63</td>
<td>32</td>
</tr>
<tr>
<td>po</td>
<td>0.00^a</td>
<td>0</td>
</tr>
<tr>
<td>ri</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>ro</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>to</td>
<td>0.00^a</td>
<td>0</td>
</tr>
<tr>
<td>prt</td>
<td>0.00^a</td>
<td>0</td>
</tr>
<tr>
<td>pro</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>pto</td>
<td>1.16</td>
<td>59</td>
</tr>
<tr>
<td>rto</td>
<td>0.00^a</td>
<td>0</td>
</tr>
<tr>
<td>prto,c</td>
<td>0.08</td>
<td>4</td>
</tr>
<tr>
<td>( \beta^2 )</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>( \phi )</td>
<td>0.04</td>
<td></td>
</tr>
</tbody>
</table>

^aA negative variance component was set to zero.
variance was the person × task interaction (32% of the total variability). Students' mean performance scores across raters and occasions depended on the particular task sampled. Some students successfully completed the Paper Towels task but performed less well on the Bugs task; vice versa for other students. Consistent with findings in military job performance (Wigdor & Green, 1991), mathematics performance (Lane, Stone, Ankenmann, & Lui, 1992), and writing achievement (e.g., Dunbar et al., 1991), large numbers of tasks may be needed to get a dependable measure of performance.

The magnitude of the other estimated sources of error was negligible, and some components were negative. Negative estimates can arise from sampling error, when the true value of the component is close to or equal to zero, or from a misspecification of the measurement model (Shavelson, Webb, & Rowley, 1989). In the present study, negative variance component estimates were due to very little variability among the means of the conditions in each facet. The means averaging across two raters and two occasions were 4.43 and 4.49 for Paper Towels and Bugs, respectively; the means for each rater averaging across two occasions and two tasks were 4.48 and 4.44. Following Brennan (1991), all negative variance component estimates were set to zero.

The generalizability coefficients for relative ($\tilde{g}^2$) and absolute ($\hat{g}$) decisions across raters, occasions, and tasks were relatively low if only one task, rater, and occasion were sampled to form the measurement (.04). Such low generalizability was due to (a) the homogeneous sample of students who, on average, scored quite high on the tasks, thereby restricting the range of scores, and (b) large measurement error attributable to the person × task and person × task × occasion interactions.

We next examined whether the magnitude of the effect of each measurement facet (raters and tasks) and combinations of these facets were consistent across different assessments and subject domains. To this end, three person × rater × task G studies were carried out (Figure 1). The person × task interaction was consistently the major source of measurement error accounting for 82%, 49%, and 48% of the total variability for the Science, Math, and CAP-Science data, respectively. Once again, task-sampling variability was the major source of error in the performance assessments. The variance components for rater, person × rater, and task × rater interactions were either zero or negligible. In other words, sampling variability due to raters does not appreciably increase measurement error or decrease generalizability. Therefore, one well-trained rater is likely to be sufficient to score most performance assessments in mathematics and science.

The relative G coefficients were 0.15, 0.21, and 0.32 for the Science, Math, and the CAP studies, respectively, when only one rater and one task were used. The absolute G coefficients were the same or only slightly lower (.15, .18, and .29, respectively), indicating the consistency across raters in scoring performance. To reach an approximate .80 relative G coefficient, about 23 tasks would be needed for the Science study, 15 tasks for the Math study, and only 8
tasks for the CAP study. About 23, 20, and 10 tasks would be needed to reach a .80 absolute G coefficient in those studies, respectively.

**School level.** Large-scale assessments report not only individual student scores but also mean scores for schools and districts. Hence, we would like to know: (a) the main sources of measurement error when schools are the objects of the measurement, (b) the generalizability of an assessment in evaluating school-level achievement, (c) the number of students to be sampled per school and the number of raters and tasks per student needed to get generalizable measures, and (d) the consistency of the magnitude of the effect of each facet at the individual and school level.

Following Cardinet, Tourneur, and Allal (1976, 1981), we carried out a person:school × rater × task G study using CAP data (see Table 2). Recall that three raters scored the performance of eight students within each of 15 schools on five science tasks. The major source of measurement error was the person: school × task interaction, accounting for 41% of the total variability in performance (see Table 4). This result, like those reported above at the individual level, indicates that students' performances were inconsistent across the sample of different science tasks; some tasks were easy to do for some students but not for other students.

Variance due to persons (students) is considered measurement error when schools are the objects of measurement. The magnitude of this variance component overshadows the remaining sources of error in the present study,
Shavelson, Baxter, and Gao

TABLE 4
Variance Component Estimates for the Person:School x Rater x Task G Study Using the CAP Data

<table>
<thead>
<tr>
<th>Source of Variability</th>
<th>Estimated Variance Component</th>
<th>Percent Total Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>School (s)</td>
<td>0.07</td>
<td>7</td>
</tr>
<tr>
<td>Rater (r)</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>Task (t)</td>
<td>0.09</td>
<td>9</td>
</tr>
<tr>
<td>Person:School (p:s)</td>
<td>0.23</td>
<td>22</td>
</tr>
<tr>
<td>st</td>
<td>0.07</td>
<td>7</td>
</tr>
<tr>
<td>rt</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>(p:s):r</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>(p:s):t</td>
<td>0.43</td>
<td>41</td>
</tr>
<tr>
<td>srt</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>(p:s):rt,e</td>
<td>0.14</td>
<td>13</td>
</tr>
</tbody>
</table>

\( \hat{\beta}^2 \) 0.08

\( \hat{\delta} \) 0.07

A negative variance component was set to zero.

accounting for 22% of the total variability. Variation among students within a school was much larger than systematic variation among schools, which accounted for only 7% of the total variability. This finding is consistent with the class-level analysis of Shavelson, Gao, and Baxter (1992).

The estimated variance due to tasks and the task \( \times \) school interaction accounted for 9% and 7% of the total variability, respectively. Some tasks were more difficult than others across all schools. Furthermore, the average performance scores of some schools on certain tasks were higher than on other tasks.

The variance components for rater, school \( \times \) rater, task \( \times \) rater, and person:school \( \times \) rater rounded to zero. The school \( \times \) rater \( \times \) task interaction accounted for about 1% of the total variability. These findings demonstrate that sampling variability due to raters is not a problem for school-level assessments but task-sampling variability is.

Consequently, we examined the effects of increasing the numbers of tasks on the generalizability of the measure with one rater in a series of decision (D) study designs. Due to the large variability among the students within a school and the large person:school \( \times \) task interaction, increasing the numbers of students sampled within a school and/or the numbers of tasks produced higher generalizability coefficients for both relative and absolute decisions (Figure 2). To reach generalizability of approximately .80 in estimating a school’s mean science achievement regardless of other schools’ performances (i.e., absolute decision), a sample of about 50 students within a school would need to be tested on 15 tasks, or about 100 students would need to be tested on 12 tasks. For rank ordering schools, however, only 25 students within a school and 10 tasks, or 100 students and 5 tasks, would be needed to reach .80 generalizabil-
Sampling Variability

FIGURE 2. *Trade-offs between numbers of tasks and students needed to achieve various levels of (a) relative and (b) absolute generalizability in CAP-Science data*

Validity

The validity of performance assessments—specifically, the convergent validity of measurement methods—was addressed within the context of Kane's
Shavelson, Baxter, and Gao

(1982) extension of G theory. The questions raised were: (a) To what extent do the achievement estimates for individual students depend on the particular tasks and/or methods sampled? and (b) do measurement methods converge in assessing students' science achievement?

One hundred eighty-six students were tested on two tasks (Electric Mysteries and Bugs) by each of four methods (observed, notebook, computer, and short answer). An examination of the variance component estimates provided in Table 5 for the $p \times t \times m$ G study using the Science data indicated that the residual term ($p \times m \times t, e$) accounted for the largest portion of the variability in performance scores (29%). Some students were more successful at the Bugs task when using observed scores but more successful at the Electric Mysteries task when using notebooks. However, the magnitude of the effect was confounded with other sources of error not explicitly controlled in the study ($e$).

The variance component for the person $\times$ task interaction accounted for 16% of the total variability: A particular student's performance (averaging over all methods) depended on the particular task. Task variability accounted for 16% of the total variability, reflecting the relative difficulty of the tasks; averaging across students and methods, the Bugs task was easier than the Electric Mysteries task.

The method effect accounted for 16% of the total variability in scores. In general, students performed best, on average across tasks, when they used computer simulations (3.90); they had lower scores on the short-answer questions (1.85) than on the other two methods (3.35 for observed and 3.21 for notebooks).

For the particular set of tasks (Electric Mysteries and Bugs), the average convergent validity coefficient between any pair of randomly sampled methods is .42. This convergent validity coefficient includes a high correlation between

<table>
<thead>
<tr>
<th>Source of Variability</th>
<th>n</th>
<th>Estimated Variance Component</th>
<th>Percent Total Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person (p)</td>
<td>186</td>
<td>0.84</td>
<td>19</td>
</tr>
<tr>
<td>Task (t)</td>
<td>2</td>
<td>0.70</td>
<td>16</td>
</tr>
<tr>
<td>Method (m)</td>
<td>4</td>
<td>0.70</td>
<td>16</td>
</tr>
<tr>
<td>$pt$</td>
<td></td>
<td>0.70</td>
<td>16</td>
</tr>
<tr>
<td>$pm$</td>
<td></td>
<td>0.14</td>
<td>3</td>
</tr>
<tr>
<td>$tm$</td>
<td></td>
<td>0.12</td>
<td>3</td>
</tr>
<tr>
<td>$ptm, e$</td>
<td></td>
<td>1.30</td>
<td>29</td>
</tr>
<tr>
<td>$\beta^2$</td>
<td></td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>$\phi$</td>
<td></td>
<td>0.19</td>
<td></td>
</tr>
</tbody>
</table>

TABLE 5
Exchangeability of Methods (Observed, Notebook, Computer, Short-Answer) and Tasks (Electric Mysteries and Bugs)

228
Sampling Variability

TABLE 6
Pearson Correlation Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Notebook</th>
<th>Computer</th>
<th>Short-Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EM Bugs</td>
<td>EM Bugs</td>
<td>EM Bugs</td>
<td>EM Bugs</td>
</tr>
<tr>
<td><strong>Observed</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electric Mysteries</td>
<td>.46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(EM)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bugs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Notebook</strong></td>
<td>.84</td>
<td>.34</td>
<td>.36</td>
<td>.71</td>
</tr>
<tr>
<td>Electric Mysteries</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(EM)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bugs</td>
<td>.36</td>
<td>.71</td>
<td>.26</td>
<td></td>
</tr>
<tr>
<td><strong>Computer</strong></td>
<td>.55</td>
<td>.46</td>
<td>.52</td>
<td>.33</td>
</tr>
<tr>
<td>Electric Mysteries</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(EM)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bugs</td>
<td>.30</td>
<td>.44</td>
<td>.22</td>
<td>.49</td>
</tr>
<tr>
<td><strong>Short-Answer</strong></td>
<td>.53</td>
<td>.35</td>
<td>.48</td>
<td>.27</td>
</tr>
<tr>
<td>Electric Mysteries</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(EM)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bugs</td>
<td>.30</td>
<td>.38</td>
<td>.22</td>
<td>.30</td>
</tr>
<tr>
<td></td>
<td>.31</td>
<td>.35</td>
<td>.31</td>
<td>.31</td>
</tr>
</tbody>
</table>

Direct observation and notebooks: $r = .84$ for Electric Mysteries and .71 for Bugs; it includes low to moderate correlations between these two methods and computer simulation and short-answer methods (see Table 6). These findings may be interpreted as indicating that not all methods converge with one another. Rather, the evidence is that certain methods may measure different aspects of achievement (cf. the $p \times m, e$ residual).

Conclusions

Development and widespread use of performance-based assessments have not, for the most part, been accompanied by systematic evaluation of their technical qualities. In this article, we bring evidence to bear on the generalizability and convergent validity of performance assessments in elementary mathematics and science. By viewing these assessments within a sampling framework, generalizability theory is used to: (a) estimate potential sources of measurement error or lack of convergence of measurement methods, (b) calculate the generalizability of the measurement, and (c) project alternative designs for collecting large-scale assessment data.

The finding that measurement error is introduced by task-sampling variability, and not by variability due to other measurement facets, is consistent with those reported elsewhere for writing achievement and military job performance. Regardless of the subject matter (mathematics or science) or the level of analysis (individual or school), large numbers of tasks are needed to get a generalizable measure of achievement. One practical implication of these findings is that—assuming 15 minutes per CAP task, for example—a total of 2.5 hours of testing time would be needed to obtain a generalizable measure (.80) of student achievement.
With regard to convergent validity, results indicate that, at least for the data reported here, student performance is dependent on methods sampled. Methods do not converge.

Findings of substantial task and method sampling variability, if replicated in further research, have important consequences for the future of performance assessments on a large scale basis. Generalizations of student or school achievement from a small sample of tasks given by one method to the domain that is defined by all tasks, raters, occasions, and methods are not supported by the data presented here.

Note

1 The difference in the magnitude of the person \( \times \) task interaction in the Science (82%) and CAP-Science (48%) studies, both involving science assessments, may be due to differences in tasks. CAP-Science tasks prescribed steps in carrying out the tasks—not so the Science study. The difference may also be due to differences in student populations. Or it may be due to a combination of these factors and some others.

References


Sampling Variability


**Authors**

RICHARD J. SHAVELSON is Professor, University of California–Santa Barbara, Graduate School of Education, Phelps Hall, Room 1198, Santa Barbara, CA 93106-9490. *Degrees*: BA, University of Oregon; MA, San Jose State University; PhD, Stanford University. *Specializations*: psychometrics, mathematics and science education reform, and policy analysis.

GAIL P. BAXTER is Assistant Professor, University of Michigan, 610 E. University, SEB 4115, Ann Arbor, MI 48109. *Degrees*: BA, Queen’s University; MA, Eastern New Mexico University; PhD, University of California–Santa Barbara. *Specialization*: educational measurement.

XIAOHONG GAO is Psychometrician, American College Testing, 2201 North Dodge, PO Box 168, Iowa City, IA 52243. *Degrees*: BA, Hangzhou University, China; MA, PhD, University of California–Santa Barbara. *Specialization*: educational measurement.
Sampling Variability of Performance Assessments
Richard J. Shavelson; Gail P. Baxter; Xiaohong Gao
Stable URL:
http://links.jstor.org/sici?sici=0022-0655%28199323%2930%3A3%3C215%3ASVOPA%3E2.0.CO%3B2-M

References

Evaluation of Procedure-Based Scoring for Hands-on Science Assessment
Gail P. Baxter; Richard J. Shavelson; Susan R. Goldman; Jerry Pine
Stable URL:
http://links.jstor.org/sici?sici=0022-0655%28199221%2929%3A1%3C1%3AEOPSFH%3E2.0.CO%3B2-2

Mathematics Performance Assessment: Technical Quality and Diverse Student Impact
Gail P. Baxter; Richard J. Shavelson; Sally J. Herman; Katharine A. Brown; James R. Valadez
Stable URL:
http://links.jstor.org/sici?sici=0021-8251%28199305%2924%3A3%3C190%3AMPATOA%3E2.0.CO%3B2-0

The Symmetry of Generalizability Theory: Applications to Educational Measurement
Jean Cardinet; Yvan Tourneur; Linda Allal
Stable URL:
http://links.jstor.org/sici?sici=0022-0655%28197622%2913%3A2%3C119%3ATSOGTA%3E2.0.CO%3B2-2

Extension of Generalizability Theory and Its Applications in Educational Measurement
Jean Cardinet; Yvan Tourneur; Linda Allal
Stable URL:
http://links.jstor.org/sici?sici=0022-0655%28198124%2918%3A4%3C183%3AEOGTAE%3E2.0.CO%3B2-G
On the Stability of Performance Assessments
Maria Araceli Ruiz-Primo; Gail P. Baxter; Richard J. Shavelson
Stable URL:
http://links.jstor.org/sici?sici=0022-0655%28199321%2930%3A1%3C41%3AOTSOPA%3E2.0.CO%3B2-H

Performance Assessments: Political Rhetoric and Measurement Reality
Richard J. Shavelson; Gail P. Baxter; Jerry Pine
Stable URL:
http://links.jstor.org/sici?sici=0013-189X%28199205%2921%3A4%3C22%3APAPRAM%3E2.0.CO%3B2-O